False Outliers

So far my Wikipedia script has churned through about 200 articles, calculating who wrote what in each. This morning I looked through them to see if there were any that didn't match my theory. It printed out a couple and I decided to investigate.

The first it found was Alkane, a long technical article about acyclic saturated hydrocarbons that it said was largely written by Physchim62. Yesterday a good friend was telling me that he thought long technical articles were likely written by a single person, so I immediately thought that here was the proof that he was right. But, just to check, I decided to look in the edit history to make sure my script hadn't made an error.

It hadn't, I found, but once again simply looking at the numbers missed the larger point. Physchim62 had indeed contributed most of the article, but according to the edit comments, it was by translating the German version! I don't have the German data, but presumably it was written in the same incremental way as most of the articles in my study.

The next serious case was Characters in Atlas Shrugged, which the script said was written by CatherineMunro. Again, it seemed plausible that one person could have written all those character bios. But again, an investigation into the actual edit history found that Munro hadn't written them, instead she'd copied them from a bunch of subpages, merging them into one bigger page.

The final serious example was Anchorage, Alaska, which appeared to have been written by JeffreyAllen1975. Here the contributions seemed quite genuine; JeffreyAllen1975 made tons of edits each contributing a paragraph at a time. The work seemed to take quite a toll on him; at his user page he noted "I just got burned-out and tired of the online encyclopedia. My time is being taken away from me by being with Wikipedia." He lasted about four months.

Still, something seemed fishy about JeffreyAllen1975, so I decided to investigate further. Currently, the Anchorage page has a tag noting that "The current version of the article or section reads like an advertisement." A bit of Googling revealed why: JeffreyAllen1975's contributions had been copied-and-pasted from other websites, like the Anchorage Chamber of Commerce ("Anchorage's public school system is ranked among the best in the nation. … The district's average SAT and ACT College entrance exam scores are consistently above the national average and Advanced Placement courses are offered at each of the district's larger high schools.").

I suspect JeffreyAllen1975 didn't know what he was doing; his writing style suggests he's just a kid: "In my free time, I am very proud of my-self by how much I've learned by making good edits on Wikipedia articles." I'm pretty sure he just thought he was helping the project: "Wikipedia is like the real encyclopedia books (A thru Z) that you see in the library, but better." But his plagiarism will still have to be removed.

When I started, just looking at the numbers these seemed to be several cases that strongly contradicted my theory. And had I just stuck to looking at the numbers, I would have believed that to be the case as well. But, once again, investigation shows the picture to be far more interesting: translation, reorganization, and plagiarism. Exciting stuff!

You should follow me on twitter here.

September 5, 2006